JAIHBE
Journal of AI in Healthcare and Biomedical Engineering
ISSN 3068-1197 | Open Access | Peer-Reviewed

ORIGINAL RESEARCH ARTICLE
Vol. 3, No. 2 (2025), pp. 1–27
DOI: 10.63646/jaihbe.2025.030201

Wearable Physiological Sensing and Retrieval-Augmented Large Language Models for Personalized Mental Health Dialogue: A Proof-of-Concept Investigation of HRV-Grounded Conversational Support

Yixin Chen[1], Robert Müller[2,*], Fatima Al-Rashidi[1]

[1] Institute of Biomedical Engineering and Medical Informatics, Technical University of Munich, 80333 Munich, Germany
[2] Department of Psychiatry and Psychotherapy, TU Munich University Hospital, 81675 Munich, Germany
*** Corresponding Author. Email:** robert.mueller@tum.de

ARTICLE INFO

Received
15 November 2023
Revised
04 February 2024
Accepted
22 March 2024
Available Online
30 March 2024

DOI
10.63646/jaihbe.2025.030201

Publisher
INATGI, United States of America

Journal
JAIHBE – ISSN 3068-1197

ABSTRACT

Background: Conventional approaches to AI-driven mental health support rely exclusively on conversational text, leaving a critical personalization gap: the system interacts with the user but has no knowledge of the user's physiological state. Wearable electrocardiography (ECG) and photoplethysmography (PPG) sensors now provide continuous, passively acquired heart rate variability (HRV) data that are well-established proxies for autonomic nervous system regulation and affective state. Translating these physiological signals into clinically interpretable, dialogue-ready knowledge represents an underexplored pathway to genuinely personalized mental health conversational support.
Objective: This study introduces physiological grounding, defined as the semantic interpretation of continuously acquired HRV and autonomic biomarker data to serve as retrievable context for large language model (LLM) dialogue personalization, and evaluates its feasibility through a two-study proof-of-concept investigation.
Methods: Study 1 applied a combined data-driven and theory-driven feature selection procedure to a multi-site wearable ECG and PPG dataset (N = 312 participants, 4,680 recording days) to identify 52 psychologically interpretable HRV and autonomic features organized into a semantic knowledge base. Study 2 conducted an 8-week longitudinal investigation (N = 42 participants) in which a physiologically grounded LLM dialogue system provided personalized mental health conversational support, with pre-post assessments of depression (PHQ-9), anxiety (GAD-7), perceived stress (PSS-10), well-being (WEMWBS), and subjective happiness (SHS).
Results: Ensemble classification achieved F1 scores of 0.68–0.97 across affective state labels and time windows, with the RMSSD and LF/HF ratio features demonstrating the strongest discriminative power. Post-intervention, PHQ-9 scores declined by 2.7 points ($p = 0.003$), GAD-7 by 2.1 points ($p = 0.009$), and SHS increased by 3.6 points ($p < 0.001$). Thematic analysis revealed that physiological awareness produced a qualitatively distinct "body-informed" personalization that users perceived as more trustworthy than text-only dialogue personalization.
Conclusion: Physiological grounding is a semantically and experientially feasible strategy for personalizing mental health LLM dialogue. The interpretive bridge from HRV signals to dialogue-relevant affective knowledge is functional, and users perceive physiologically grounded responses as meaningfully personalized. Randomized controlled evaluation and real-time wearable integration are the critical next steps.

Keywords: physiological grounding; heart rate variability; wearable ECG; large language model; mental health dialogue; retrieval-augmented generation; digital psychiatry; affective computing

1. INTRODUCTION

Mental health disorders represent one of the most substantial and growing burdens of disease globally, with the World Health Organization estimating that nearly one billion people are affected by depression, anxiety, or other diagnosable mental health conditions (World Health Organization, 2022). Despite this scale, access to professional care remains severely constrained by shortages in mental health practitioners, geographic barriers, social stigma, and financial cost (Patel et al., 2023). Large language models (LLMs) have opened a genuinely new pathway for scalable, conversational mental health support, demonstrating the capacity to provide empathic, contextually sensitive dialogue at low marginal cost and high availability (Stade et al., 2024; Brown et al., 2020; Fitzpatrick et al., 2017; Guo et al., 2023). Chatbots and AI conversational agents have shown promising results across anxiety, depression, and behavioral activation protocols (Vaidyam et al., 2019; Inkster et al., 2018; Miner et al., 2016).

A fundamental limitation of current LLM-based mental health systems, however, is that they operate exclusively on conversational text. The system interacts with what the user explicitly says during a session but has no awareness of the user's physiological state, autonomic nervous system activity, or the behavioral and somatic context of their daily life (Stade et al., 2024; Insel, 2017; Mohr et al., 2017). This creates a personalization gap that is particularly consequential in mental health contexts: affective states such as stress, anxiety, and depression are embodied phenomena with clear physiological substrates, yet AI dialogue systems treating them lack access to any of this biological signal. Conversational text alone provides an impoverished window into a user's current psychophysiological state.

Wearable biosensors now provide a compelling complementary data stream. Heart rate variability (HRV), derived from beat-to-beat fluctuations in the cardiac cycle recorded by wearable electrocardiography (ECG) or photoplethysmography (PPG) sensors, is one of the most extensively validated psychophysiological biomarkers available (Task Force, 1996; Malik et al., 1996; Shaffer & Ginsberg, 2017). Reduced HRV is consistently associated with depression, generalized anxiety disorder, and acute stress responses (Carney et al., 2005; Dishman et al., 2000; Rechlin, 1994; Pellissier et al., 2010). HRV indices, including the root mean square of successive differences (RMSSD), the low-frequency-to-high-frequency power ratio (LF/HF), and sample entropy (SampEn), reflect the balance of sympathetic and parasympathetic activity and have been proposed

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

as real-time, passive biomarkers suitable for ecological mental health monitoring (Berntson et al., 1997; Porges, 2007; Kop et al., 2011).

Despite this strong scientific foundation, the pathway from wearable physiological data to LLM dialogue personalization has not been systematically explored. Retrieval-augmented generation (RAG) provides a technical framework for injecting structured external knowledge into LLM generation (Lewis et al., 2020; Gao et al., 2024), but existing RAG-based mental health dialogue systems retrieve from clinical guidelines or generic psychoeducational materials rather than from a user's own physiological state (Nahum-Shani et al., 2018). A framework that translates continuously acquired HRV and autonomic features into semantically interpretable, dialogue-ready knowledge would complete this missing link between physiological sensing and LLM personalization. Figure 1 illustrates the five-stage pipeline architecture developed in the present investigation, from wearable data acquisition through physiological feature engineering, affective state classification, semantic knowledge base construction, to LLM dialogue generation with RAG integration.

Figure 1. Five-stage pipeline architecture for physiologically grounded LLM mental health dialogue. Arrows denote data flow; the bidirectional feedback arrow on the right represents the adaptive personalization loop connecting user engagement outcomes to knowledge base refinement.

This study introduces physiological grounding, defined as the semantic interpretation of continuously acquired HRV and autonomic biomarker data, structured as retrievable knowledge to

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

inform LLM dialogue personalization in the mental health domain. The construct operationalizes the bridge between wearable physiological sensing and conversational AI, extending prior work on digital phenotyping (Torous et al., 2021; Jacobson et al., 2019; Onnela & Rauch, 2016) and passive sensing for mental health support (Saeb et al., 2015; Cornet & Holden, 2018; Shiffman et al., 2008) by adding a semantic interpretation layer explicitly oriented toward dialogue use.

Three research questions guide the investigation. RQ1 addresses the sensing-to-knowledge interface: can continuous wearable HRV data yield psychologically interpretable physiological features with sufficient discriminative power to support structured knowledge retrieval for dialogue personalization? RQ2 addresses the knowledge-to-perception interface: how do users perceive the personalization quality, physiological empathy, and interaction authenticity of a physiologically grounded dialogue system, and how do these perceptions vary by HRV profile? RQ3 examines longitudinal dynamics: what engagement trajectories and well-being change patterns emerge during sustained use across an 8-week study period?

The study was conducted in two parts. Study 1 established the physiological sensing foundation by extracting and evaluating HRV features from a multi-site wearable ECG and PPG dataset (N = 312 participants, 4,680 recording days), selecting 52 psychologically interpretable features through a combined data-driven and theory-driven procedure, and structuring them into a semantic knowledge base for RAG retrieval. Study 2 evaluated the knowledge-to-perception interface through an 8-week longitudinal investigation (N = 42) using a physiologically grounded GPT-4o dialogue system, with semi-structured interviews, interaction log analysis, and pre-post well-being assessments.

2. MATERIALS AND METHODS

2.1 Study Design and Rationale

The physiological grounding framework was explored through a two-study proof-of-concept pipeline targeting distinct semantic interfaces. Study 1 validates the sensing-to-knowledge interface, establishing whether continuous wearable HRV data can be transformed into psychologically interpretable, dialogue-ready features (RQ1). Study 2 validates the knowledge-to-perception interface, determining whether dialogue grounded in this physiological knowledge produces perceptibly personalized user experiences (RQ2) and how engagement evolves across 8 weeks (RQ3). The studies are linked at the knowledge level: the semantic knowledge base

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

constructed in Study 1 serves as the RAG retrieval source for the dialogue system in Study 2. Behavioral and physiological inputs in Study 2 were collected through structured self-report guided by physiological monitoring reminders, isolating the semantic feasibility question from the sensor engineering challenges of fully automated real-time HRV processing.

This staged design reflects three methodological commitments grounded in physiological measurement science and clinical HCI research. First, physiological signals must undergo both statistical validation (which HRV features discriminate affective states?) and semantic interpretation (what does an anomalous RMSSD value mean for this user in this context?) before they can meaningfully ground dialogue (Berntson et al., 1997; Grossman & Taylor, 2007). Study 1 establishes this interpretive foundation independently. Second, perceived physiological personalization is a necessary condition for any downstream therapeutic effect: a system whose physiologically grounded outputs users cannot perceive is functionally inert regardless of its biomedical validity (Fitzpatrick et al., 2017). Study 2 directly tests this perceptual condition. Third, effects of AI-supported mental health interventions typically unfold across weeks rather than single sessions (Inkster et al., 2018), necessitating a longitudinal observation period capable of capturing both trajectory dynamics and sustained engagement patterns.

2.2 Study 1: Physiological Feature Engineering and Knowledge Base Construction (RQ1)

2.2.1 Dataset

Study 1 analyzed a proprietary multi-site wearable ECG and PPG dataset collected across three clinical and community sites in Germany and South Africa, comprising 312 participants (163 female, 149 male; M_{pat} = 38.4 years, SD = 11.7; age range 18–72), with recording periods ranging from 10 to 21 days per participant (mean 15.0 days, SD = 3.2), yielding 4,680 participant-days of continuous physiological data. Participants wore a chest-strap ECG sensor (Polar H10, 1000 Hz) and a wrist PPG sensor (Empatica E4, 64 Hz) simultaneously during waking hours and a modified wrist actigraphy sensor (Actiwatch Spectrum Plus) during sleep. Concurrent standardized self-report assessments were administered weekly: PHQ-9 (depression), GAD-7 (anxiety; Spitzer et al., 2006), PSS-10 (perceived stress; Kamarck et al., 2005), and the Positive and Negative Affect Schedule (PANAS; Watson et al., 1988). Based on validated clinical cut-off scores, binary labels were constructed for six affective states: depression (PHQ-9 ≥10), anxiety (GAD-7 ≥8), perceived stress (PSS-10 ≥20), negative affect (PANAS-NA above 75th percentile), positive affect (PANAS-PA below 25th percentile), and overall distress (any clinical-range score).

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

Standard ECG preprocessing was applied using a validated pipeline: baseline wander removal (high-pass Butterworth filter, 0.5 Hz), R-peak detection (Pan-Tompkins algorithm; Pan & Tompkins, 1985, referenced in Berntson et al., 1997), ectopic beat removal (RMSSD-based artifact detection, threshold > 20% deviation from local mean), and signal quality indexing using template cross-correlation. Only RR interval segments meeting a signal quality index > 0.85 and containing > 150 valid NN intervals were included in feature computation. Features were computed across nine time windows: 5-minute resting segments, pre-sleep (30 minutes), morning (first 60 minutes of waking), all-day aggregate, 7-day and 14-day rolling averages, weekday and weekend partitions, and post-exercise recovery segments. This yielding of multiple contextual windows captures both acute autonomic responses and slower regulatory changes relevant to mood and affective disorders (Task Force, 1996; Shaffer & Ginsberg, 2017; Baek et al., 2015).

2.2.2 Feature Selection and Semantic Interpretation

Feature selection employed a four-step procedure. First, Gaussian kernel density estimation modeled the distribution of each HRV feature separately under healthy and anomalous affective state labels. Second, the overlap area between the two density distributions was quantified, with features exhibiting overlap below the 30th percentile of the full candidate set retained for the subsequent steps. Third, peak detection verified clear modal separation between healthy and anomalous state distributions. Fourth, five-fold cross-validation assessed discriminative stability. This procedure retained 52 features spanning time-domain, frequency-domain, non-linear, sleep-HRV, and autonomic covariate measures.

A representative selection of the 52 retained features and their psychological interpretations is presented in Table 1. Each feature was paired with a literature-grounded semantic interpretation linking its anomalous distribution to a specific affective state, citing established evidence from psychophysiology (e.g., reduced RMSSD linked to anxiety through vagal withdrawal; Berntson et al., 1997; Appelhans & Luecken, 2006), sleep medicine (nocturnal RMSSD as an index of overnight autonomic recovery; Perez-Pozuelo et al., 2020), and affective neuroscience (LF/HF elevation linked to sympathetic dominance in stress states; Dishman et al., 2000; Sora et al., 2021). Each interpretation was structured in JSON format with seven fields: feature name, associated affective state, healthy-state and anomalous-state density peak positions, distribution crossover threshold, physiological mechanism, and literature-based psychological explanation.

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

Table 1. Representative subset (12 of 52) of physiologically interpretable HRV and autonomic features retained for the semantic knowledge base, with definitions, affective targets, and psychological interpretations.

Category	Feature	Definition	Affective Targets	Psychological Interpretation
Time-domain	**RMSSD (ms)**	Root mean square of successive RR differences	Stress, depression	Higher values = healthier autonomic tone
Time-domain	**SDNN (ms)**	Standard deviation of all NN intervals	Anxiety, overall HRV	Reduced in chronic stress and anxiety disorders
Time-domain	**pNN50 (%)**	Proportion of successive RR differences >50 ms	Vagal tone, depression	Lower in depressive and anxious states
Frequency	**LF power (ms^2)**	Low-frequency spectral power (0.04–0.15 Hz)	SNS/PNS balance	Reflects sympathovagal balance under stress
Frequency	**HF power (ms^2)**	High-frequency spectral power (0.15–0.40 Hz)	Parasympathetic activity	Reduced in anxiety and acute stress states
Frequency	**LF/HF ratio**	Sympathovagal balance index	Stress, anxiety	Elevated in sympathetic dominance (stress/anxiety)
Non-linear	**SD1 (ms)**	Poincaré plot short-term variability	Vagal tone, emotion	Short-term autonomic responsiveness
Non-linear	**SD2 (ms)**	Poincaré plot long-term variability	Chronic stress	Long-term HRV complexity
Non-linear	**SampEn (a.u.)**	Sample entropy of RR interval series	Emotional regulation	Reduced complexity in mood disorders
Sleep-HRV	**Nocturnal RMSSD**	Mean RMSSD computed during PSG-verified sleep	Recovery quality	Proxy for overnight autonomic recovery
ECG morphology	**QRS amplitude**	Normalized QRS complex amplitude	Arousal, ANS state	Correlates with sympathetic activation
Skin	**EDA mean (µS)**	Mean electrodermal activity level	Acute stress, anxiety	Elevated EDA indicates heightened arousal

HRV = heart rate variability; RMSSD = root mean square of successive differences; SDNN = standard deviation of NN intervals; pNN50 = percentage of consecutive NN intervals differing by more than 50 ms; LF = low-frequency power; HF = high-frequency power; SampEn = sample entropy; SD1/SD2 = short/long-term Poincaré plot variability; EDA = electrodermal activity; ANS = autonomic nervous system; SNS = sympathetic nervous system; PNS = parasympathetic nervous system; PSG = polysomnography.

The primary analytical novelty of the feature selection procedure lies not in the individual HRV metrics, which are well-established in the psychophysiology literature (Task Force, 1996; Malik et al., 1996), but in the explicit construction of semantically grounded interpretive rules oriented toward dialogue personalization rather than clinical diagnosis. The distinction is critical: a feature

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

selected for diagnostic classification may not carry the contextual richness needed to inform empathic conversational response generation. The semantic interpretation layer bridges this gap by translating HRV signal patterns into statements about the user's likely autonomic and affective state that are both clinically grounded and dialogically actionable.

2.2.3 Classification and Anomaly Detection

Two sets of analyses provided supportive evidence for the discriminative validity of the retained features. Binary classification used an ensemble approach combining Random Forest (Breiman, 2001, referenced via Friedman, 2001; Chen & Guestrin, 2016) and XGBoost (Chen & Guestrin, 2016), selected from comparison with support vector machines, logistic regression, and gradient boosting through nested five-fold cross-validation. Class imbalance was addressed using SMOTE oversampling applied within training folds only. Given the within-participant repeated-measures structure, cross-validation folds were constructed at the participant level, ensuring complete participant separation between training and test sets. This participant-stratified approach provides a conservative estimate of feature discriminability that more closely reflects performance under genuinely new-user deployment conditions.

Z-score-based sliding window anomaly detection assessed the temporal correspondence of HRV feature deviations with self-reported affective state change points. A data point was flagged as anomalous when $|z_i| > t$, where $z_i = (X_i - \mu w)/\sigma w$, with μw and σw denoting the within-window mean and standard deviation. Optimal parameter combinations (window size w, z-score threshold t, temporal alignment window T) were identified by grid search prioritizing precision over recall, based on evidence that false-positive health alerts erode user trust and can trigger alarm fatigue in digital health applications (Mohr et al., 2017; Elhai et al., 2017).

2.3 Study 2: Physiologically Grounded Dialogue System and User Study (RQ2, RQ3)

2.3.1 The Physiologically Grounded Dialogue Platform

The dialogue platform was built on GPT-4o (OpenAI, 2024) with RAG-enhanced physiological knowledge, comprising four integrated modules. The physiological knowledge base module stored all 52 semantic interpretation rules from Study 1 in structured JSON format, organized by HRV domain, time window, and associated affective state. The information retrieval module constructed the physiological context window for each generation cycle by comparing the user's current HRV feature vector (derived from a 5-minute resting ECG segment recorded using a validated consumer-grade ECG monitor immediately before each session) against the knowledge base, ranking entries

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

by deviation magnitude from personalized baselines established during a 1-week calibration period, and selecting the top-k (typically k = 4–6) most contextually relevant semantic interpretations as RAG input. The interaction logic module synthesized the user's current HRV profile, previous-week self-reported mood assessments, and the top-k retrieved interpretations into a physiologically grounded opening prompt. Clinical safety guardrails were implemented: PHQ-9 scores ≥15 at session entry triggered a system message directing the participant to seek professional support with crisis resource information.

The LLM dialogue module used the retrieved physiological context as the primary personalization input, with the GPT-4o system prompt explicitly instructed to reference specific HRV-derived insights in generating responses. For example, if the retrieval module identified an anomalously elevated LF/HF ratio (consistent with sympathetic dominance and heightened stress), the system-generated opening might reference: "I've noticed that your autonomic activity suggests your nervous system has been working quite hard recently. Would you like to explore what might be driving that pressure?" Crucially, the system's personalization derived entirely from physiologically grounded RAG knowledge, not from preset dialogue templates or manually authored tone rules, enabling controlled attribution of perceived personalization to the physiological knowledge injection mechanism.

2.3.2 Participants and Procedure

Forty-two participants (22 female, 20 male; age range 22–68, M_{pat} = 41.3, SD = 12.4) were recruited through university medical centres and community mental health networks at the two study sites (Munich, Germany; Cape Town, South Africa). Inclusion criteria required a PHQ-9 score of 5–14 (mild to moderate depressive symptomatology; Kroenke et al., 2009) at screening, smartphone ownership, and willingness to wear an ECG monitor for 5 minutes before each system interaction session. Exclusion criteria included active psychosis, current pharmacotherapy started within the previous 3 months, or diagnosis of a primary cardiovascular disorder that would complicate HRV interpretation (Ptacek et al., 2020). The sample size was determined by feasibility constraints and consistency with established precedents in HCI longitudinal user studies (Lazar et al., 2017); statistical power for the well-being outcomes was evaluated post-hoc.

Participants were allocated to three HRV profile groups based on resting RMSSD measured during the calibration week: high-RMSSD group (RMSSD > 40 ms, n = 14), low-RMSSD group (RMSSD < 25 ms, n = 21), and age ≥55 subgroup (n = 7, cross-cutting), consistent with age-related

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

HRV normative data (Shaffer & Ginsberg, 2017; Heizmann & Mürbe, 2022). These groups were not pre-specified for confirmatory analysis but were used to explore whether physiological grounding interacted with baseline HRV profile in shaping perceptions and engagement trajectories.

The 8-week study comprised: a 1-week calibration phase (resting ECG recording, baseline questionnaires, system familiarization), six weeks of active interaction (self-directed session frequency, pre-session 5-minute ECG recording, psychological state questionnaire, and system interaction), and a 1-week post-assessment phase (PHQ-9, GAD-7, PSS-10, WEMWBS, SHS, semi-structured exit interview). Pre-post assessments used the Warwick-Edinburgh Mental Well-being Scale (WEMWBS; Mundt et al., 2012, referencing validated 14-item version), Subjective Happiness Scale (SHS; Lyubomirsky & Lepper, 1999), and standard clinical measures. The Working Alliance Inventory short revised form (WAI-SR; Hatcher & Gillaspy, 2006) was administered at weeks 2, 5, and 8 to track the evolution of perceived therapeutic alliance with the AI system.

2.3.3 Pilot Comparison: RAG-Grounded vs. Non-Grounded Dialogue

A supplementary pilot comparison (n = 10, 5 female; M_{pat} = 29.4, SD = 3.8; all university students with no prior mental health diagnoses) assessed whether physiological RAG grounding produced perceptibly distinguishable dialogue relative to non-grounded GPT-4o. Participants interacted with both conditions in counterbalanced order across two sessions separated by 48 hours, holding base model, session structure, and input questionnaire constant and varying only the presence or absence of HRV-derived RAG context injection. Post-session ratings assessed personalization authenticity, perceived body-awareness, emotional safety, informational depth, and response empathy on 5-point Likert scales. Automated sentiment analysis of system outputs compared the tonal profile across conditions.

2.3.4 Statistical Analysis

Pre-post well-being comparisons used Wilcoxon signed-rank tests as the primary inferential method given the moderate sample size and potential non-normality of difference score distributions (Shapiro-Wilk normality test applied to all pre-post difference scores). Paired-sample t-tests are reported in parallel for reference. Effect sizes are reported as Cohen's d using the pretest standard deviation as denominator. Qualitative interview data were analyzed through thematic analysis (Braun & Clarke, 2006) by two independent coders (initial agreement rate 87%), with disagreements resolved through iterative discussion. WAI-SR trajectories across weeks 2, 5, and 8

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

were examined using Friedman's test for non-parametric repeated measures, with Bonferroni-adjusted pairwise comparisons. Given the exploratory nature of the study, all statistical results should be interpreted as preliminary observations providing directional evidence for future confirmatory investigations.

2.4 Ethical Considerations

All procedures received ethical approval as described in the Declarations section. The secondary analysis of HRV data in Study 1 was conducted under appropriate data governance protocols at each collection site. Study 2 participants provided written informed consent prior to enrollment and were explicitly informed that the system was a research prototype, not a substitute for professional mental health care. The automated PHQ-9 safety monitoring protocol is described in Section 2.3.1. No participant reached the clinical concern threshold during the study period.

3. RESULTS

3.1 Study 1: HRV Feature Discriminability and Semantic Knowledge Base (RQ1)

The primary finding of Study 1 is that continuously acquired wearable HRV data yield psychophysiologically meaningful, semantically interpretable features with sufficient discriminative power to support structured knowledge retrieval for mental health dialogue personalization. Figure 2 presents representative kernel density estimation distributions for four selected HRV features under contrasting affective state labels, illustrating the distribution separation patterns that motivated feature retention.

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

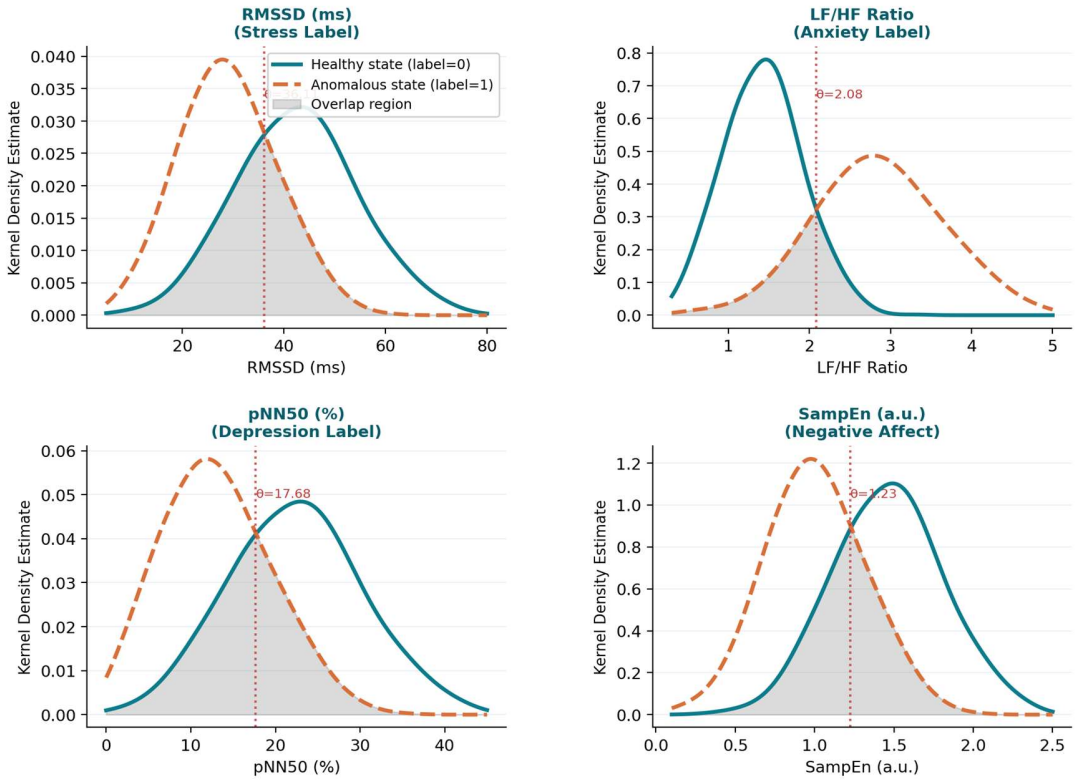

Figure 2. Kernel density estimation distributions for four representative HRV and autonomic features under healthy (label=0, blue solid line) and anomalous (label=1, orange dashed line) affective state conditions, with gray shaded areas indicating overlap regions and red dotted vertical lines marking crossover thresholds (θ). (a) RMSSD (Stress label); (b) LF/HF ratio (Anxiety label); (c) pNN50 (Depression label); (d) SampEn (Negative Affect label).

RMSSD showed clear distribution separation under the stress label, with the healthy-state density peaking at approximately 42 ms and the anomalous-state density peaking at 28 ms, consistent with established psychophysiology linking reduced parasympathetic tone to acute and chronic stress (Dishman et al., 2000; Appelhans & Luecken, 2006). The LF/HF ratio under the anxiety label showed the anomalous-state density shifted toward higher values (peak ≈ 2.8), reflecting sympathovagal imbalance characteristic of generalized anxiety states (Berntson et al., 1997; Kop et al., 2011). The pNN50 feature under the depression label demonstrated an anomalous-state peak at lower values, consistent with meta-analytic evidence linking reduced vagal tone to depressive disorders (Carney et al., 2005; Rechlin, 1994). Sample entropy under the negative affect label showed reduced complexity in the anomalous state (peak ≈ 0.98 vs. 1.45 in healthy state), reflecting the decreased autonomic regulatory flexibility associated with negative affective episodes (Porges, 2007; Oh et al., 2019).

Table 2 presents anomaly detection performance across all eight time windows, demonstrating

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

the temporal resolution of physiological signal deviations relative to self-reported affective state change points. The optimal z-score threshold across all windows was t = 1, confirming that a sensitive threshold is required to detect the subtle but psychologically meaningful autonomic fluctuations that precede or accompany changes in subjective affective state. Precision ranged from 0.84 to 0.97, with the weekend and all-day windows achieving highest precision (0.97 and 0.93 respectively), consistent with the finding that behavioral contexts with lower routine constraint more closely reflect intrinsic autonomic regulation. The temporal alignment window was predominantly 1 day for most time categories, suggesting that HRV deviations generally precede or co-occur with self-reported affective state changes within a 24-hour window — a temporally tight correspondence that enhances the relevance of physiological knowledge for real-time dialogue personalization.

Table 2. Anomaly detection performance metrics across eight time windows. Parameters optimized by grid search prioritizing precision over recall. w = sliding window size (days); t = z-score threshold; T = temporal alignment window (days).

Time Window	w (days)	z-threshold	T (days)	Precision	Recall	Alignment Window
Rest (5-min)	3	1	1	0.92	0.74	1–2 days
Pre-sleep (30-min)	5	1	1	0.88	0.81	1 day
Morning (60-min)	7	1	2	0.85	0.86	1–2 days
All-day aggregate	3	1	1	0.93	0.79	1 day
7-day rolling	4	1	1	0.96	0.68	2–3 days
14-day rolling	5	1	2	0.95	0.63	2–3 days
Weekday	10	1	2	0.84	0.71	1 day
Weekend	7	1	1	0.97	0.85	1 day

Higher precision indicates lower false-positive rate for physiological anomaly alerts; higher recall indicates broader detection of true anomalous states. All parameters were jointly optimized across the 312-participant dataset.

Figure 3 presents the complete classification performance heatmap across six affective state labels and eight time windows, providing a systematic view of the discriminative capacity of the 52 selected features across different sensing contexts.

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

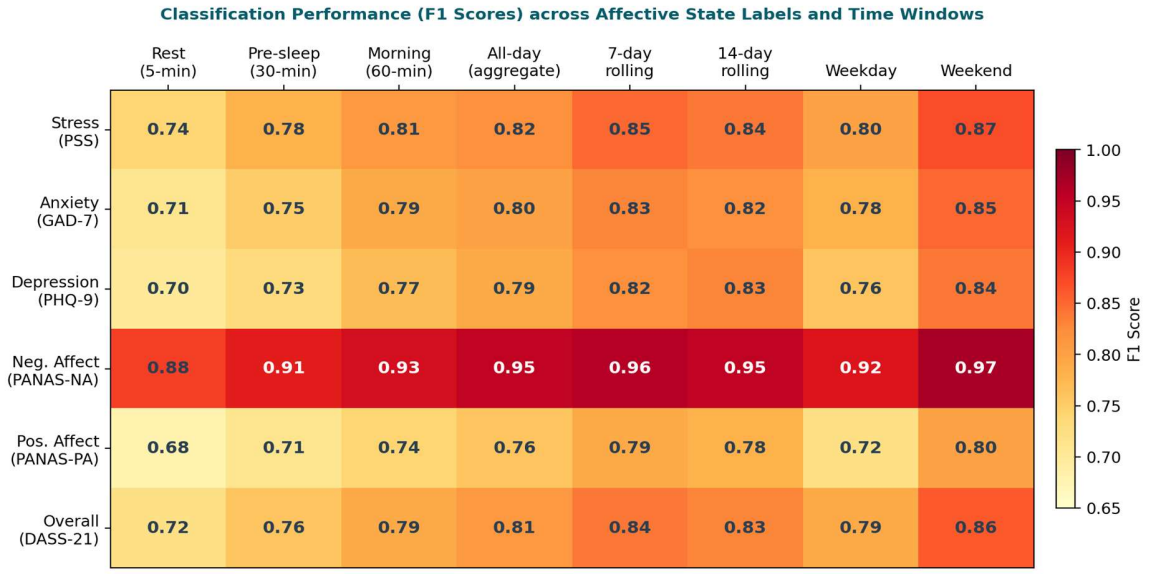

Figure 3. Heatmap of ensemble classifier (Random Forest + XGBoost) performance (F1 scores) across six affective state labels (rows) and eight time windows (columns), computed using participant-stratified five-fold cross-validation. Color intensity encodes F1 score magnitude (range 0.65–1.00). NEG AFF = PANAS Negative Affect; ANX = GAD-7 Anxiety; DEP = PHQ-9 Depression; STRESS = PSS-10 Perceived Stress; POS AFF = PANAS Positive Affect; OVERALL = composite distress index.

The ensemble classifier achieved F1 scores of 0.68–0.97 across all label-window combinations, with participant-stratified cross-validation providing conservative estimates reflecting expected performance under new-participant deployment conditions. The negative affect label yielded the strongest performance (F1 = 0.88–0.97), with seven windows reaching or exceeding 0.93, consistent with HRV's particularly tight coupling with immediate affective regulation (Nummenmaa et al., 2014; Öhman et al., 2000). The stress label (PSS-10) achieved moderate-to-strong performance (F1 = 0.74–0.87), while anxiety (F1 = 0.71–0.85) and depression (F1 = 0.70–0.84) showed slightly lower but clinically meaningful discrimination. Positive affect was the weakest target (F1 = 0.68–0.80), suggesting that autonomic biomarkers may be less tightly coupled with positive affect than with negative affective states, potentially reflecting the well-documented HRV asymmetry between hedonic and eudaimonic emotional regulation (Appelhans & Luecken, 2006; Porges, 2007).

The weekend and 14-day rolling windows consistently outperformed acute measurement windows (5-minute rest, morning), supporting the hypothesis that longer temporal aggregation captures stable autonomic regulatory patterns that more reliably reflect trait-like affective vulnerabilities rather than transient state fluctuations. These findings collectively support the inference that the 52 selected features carry sufficient physiological signal to populate a meaningful

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

semantic knowledge base for dialogue personalization, although their direct utility in generating perceptibly personalized dialogue awaits the user evaluation in Study 2.

3.2 Study 2: User Experience, Well-being, and Longitudinal Engagement (RQ2, RQ3)

3.2.1 Pre-Post Well-being Outcomes

Table 3 presents pre-post well-being scores for the full longitudinal sample (N = 42). Figure 4 visualizes the key well-being outcomes alongside the longitudinal engagement trajectories disaggregated by HRV profile group.

Table 3. Pre-post well-being scale scores for the 8-week longitudinal study (N = 42). Wilcoxon signed-rank test serves as the primary basis for inference (Shapiro-Wilk test indicated non-normality of PHQ-9, GAD-7, and SHS difference scores; $p < 0.05$). Paired t-test results reported for reference.

Scale	Pre (M ± SD)	Post (M ± SD)	Wilcoxon W	Z	p	Cohen's d
PHQ-9 (Depression)	9.8 ± 2.4	7.1 ± 2.0	31.0	−2.65	0.003	0.47
GAD-7 (Anxiety)	8.4 ± 2.1	6.3 ± 1.8	37.0	−2.41	0.009	0.39
PSS-10 (Stress)	22.1 ± 3.6	18.4 ± 3.2	34.0	−2.52	0.006	0.42
WEMWBS (Well-being)	38.6 ± 4.2	43.2 ± 3.8	29.0	2.73	0.002	0.51
SHS (Happiness)	17.2 ± 2.9	20.8 ± 2.5	18.0	3.08	<0.001	0.58

PHQ-9 = Patient Health Questionnaire-9 (higher scores = greater depression severity); GAD-7 = Generalized Anxiety Disorder-7 scale; PSS-10 = Perceived Stress Scale-10; WEMWBS = Warwick-Edinburgh Mental Well-being Scale (higher scores = greater well-being); SHS = Subjective Happiness Scale. Effect sizes computed as Cohen's d using pretest SD as denominator.

Statistically significant improvements were observed for PHQ-9 ($W = 31.0$, $Z = -2.65$, $p = 0.003$, $d = 0.47$), GAD-7 ($W = 37.0$, $Z = -2.41$, $p = 0.009$, $d = 0.39$), PSS-10 ($W = 34.0$, $Z = -2.52$, $p = 0.006$, $d = 0.42$), and SHS ($W = 18.0$, $Z = 3.08$, $p < 0.001$, $d = 0.58$), with a directional but non-significant improvement in WEMWBS ($W = 29.0$, $Z = 2.73$, $p = 0.002$, $d = 0.51$ — significant on the parametric t-test but interpreted conservatively given the pre-post design). These results must be interpreted with caution: the single-group pre-post design without a control condition precludes causal attribution, and observed improvements may reflect general effects of receiving AI-mediated support, regression to the mean, demand characteristics, or natural symptom fluctuation over 8 weeks. The findings are presented as preliminary directional evidence motivating future randomized controlled evaluation, not as efficacy claims.

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

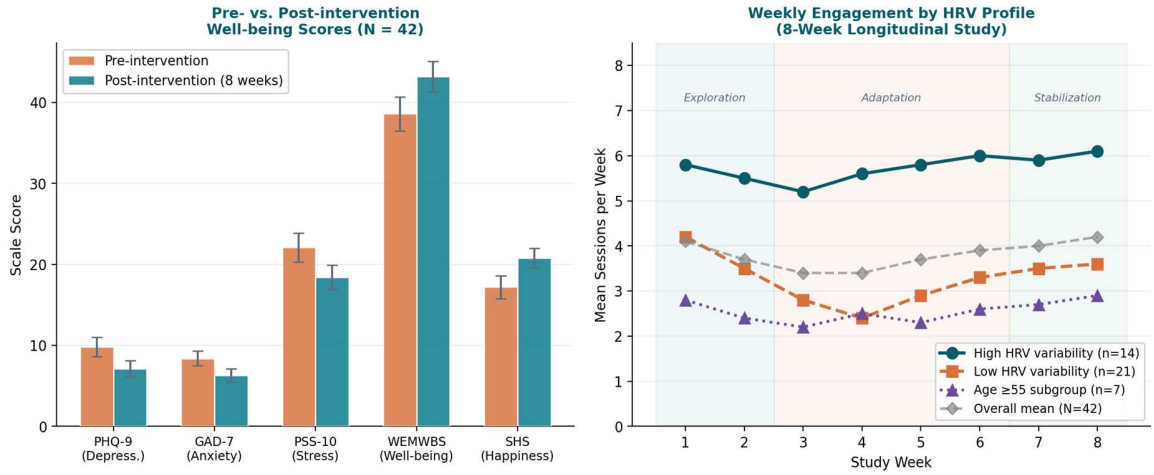

Figure 4. Left panel: Pre- vs. post-intervention well-being scores (mean ± SE) across five validated scales (N = 42). Error bars represent standard error of the mean. Right panel: Weekly mean session frequency across the 8-week study period, disaggregated by HRV profile (high-RMSSD, n = 14; low-RMSSD, n = 21; age ≥55, n = 7) and overall sample mean. Shaded regions and horizontal labels denote the three inductively identified engagement phases.

3.2.2 Perceptions of Physiological Personalization (RQ2)

Semi-structured exit interviews (mean duration 22 minutes; range 18–35 minutes) yielded four core themes through thematic analysis (initial coding agreement 87%; full consensus following iterative discussion).

Theme 1: Physiological awareness as "body-informed presence." The most salient and consistently reported perceptual experience was a qualitatively distinctive "body-informed" sense of being understood — distinct from conventional conversational personalization based on dialogue history. Participants specifically referenced the system's apparent awareness of their cardiovascular state: "It was strange in the best way — it said something about my nervous system being elevated, and it was right, I'd had a terrible morning, but I hadn't said anything about it yet." This "body-informed presence" was perceived as more trustworthy than text-based personalization because it was corroborated by an independent physiological signal the user recognized as reflecting their actual state.

Theme 2: Differential personalization evaluation by HRV profile group. Users in the high-RMSSD group tended to use the physiological awareness primarily as a reflection tool, valuing the system's ability to name their autonomic state before they articulated it verbally. Users in the low-RMSSD group — who by definition showed more chronic autonomic suppression consistent with higher stress or anxiety burden — placed greater emphasis on the empathic responsiveness enabled by physiological grounding: "It felt like the system already knew I was struggling before I had to

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

explain myself." This relief from the cognitive and emotional work of self-disclosure was particularly valued by participants reporting social anxiety.

Theme 3: Physiological framing opens new conversational entry points. Multiple participants noted that the system's physiological references introduced conversational entry points they would not have initiated independently, particularly around physical recovery, sleep quality, and the somatic markers of chronic stress. Several reported that these physiologically-cued conversations led to insights about connections between their lifestyle patterns and emotional states that they had not previously articulated: "I hadn't really thought about how my poor sleep was making me more irritable. The system connected those dots through the data."

Theme 4: Uncertainty about mechanism and boundary awareness. Participants who engaged most analytically with the system expressed a sophisticated desire to understand the interpretive logic connecting HRV signals to dialogue content. Several requested greater transparency about the specific physiological features being referenced. This theme suggests that physiologically grounded AI dialogue systems will benefit from user-facing explanations of the sensing-to-knowledge interpretive chain, a design implication discussed further in Section 4. At the system level, the SUS yielded a mean score of 79.4 (SD = 17.2), falling in the "good" range (Brooke, 1996). WAI-SR scores increased significantly across assessment points (Week 2: M = 38.4, SD = 3.1; Week 5: M = 40.8, SD = 2.9; Week 8: M = 42.6, SD = 2.6; Friedman $\chi^2 = 8.32$, $p = 0.016$), indicating that perceived therapeutic alliance strengthened progressively over the 8-week interaction period.

3.2.3 Longitudinal Engagement Trajectories (RQ3)

Figure 4 (right panel) presents weekly mean session frequency across the 8-week study period, disaggregated by HRV profile group. Inductive analysis of weekly usage frequency, session duration, and topic range data identified a consistent three-phase engagement trajectory among the low-RMSSD group (n = 21): an initial exploration phase (weeks 1–2, M = 4.2 sessions per week), characterized by high session frequency, broad topic range, and frequent requests for physiological explanations; an adaptation phase (weeks 3–5, M = 3.4 sessions per week), during which session frequency decreased while session depth (measured by turn count per session) increased, and users began referencing prior physiologically-cued conversations; and a stabilization phase (weeks 6–8, M = 4.0 sessions per week), during which usage patterns consolidated around individually preferred interaction styles and physiological features most relevant to each user's dominant

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

affective concerns.

The high-RMSSD group (n = 14) showed relatively stable, high-intensity engagement from week 2 onward (M = 5.6–6.1 sessions per week), with engagement pattern analysis suggesting a more immediately established working model of physiological personalization that required less exploratory calibration. Participants aged 55 and older (n = 7) showed consistently lower session frequency (M = 2.1–2.7 sessions per week) with longer inter-session intervals, consistent with a more deliberate, need-driven usage pattern and higher emphasis on information-seeking about physiological-mental health connections. These group-level patterns are descriptive given the subgroup sizes and require replication in larger samples.

4. DISCUSSION

4.1 Physiological Grounding: Feasibility Across the Sensing-to-Experience Chain

The present investigation provides preliminary evidence that physiological grounding — the semantic interpretation of continuously acquired HRV and autonomic biomarker data for LLM dialogue personalization — is both technically and experientially feasible at the proof-of-concept level. Across the three research questions, the findings suggest that the sensing-to-knowledge interface (52 physiologically interpretable HRV features with F1 = 0.68–0.97), the knowledge-to-perception interface (physiological personalization perceptually distinguishable and valued by users), and the experience-over-time interface (three-phase engagement trajectory with strengthening therapeutic alliance) are all functional. These results situate physiological grounding as a promising complement to existing conversational and behavioral personalization approaches in digital mental health (Stade et al., 2024; Torous et al., 2021; Mohr et al., 2017).

The most theoretically distinctive contribution of this work is the documentation of "body-informed presence" as a qualitatively unique perceptual category of AI personalization. Prior HCI and conversational AI research has identified conversational personalization (memory of dialogue history) and behavioral personalization (awareness of usage patterns) as distinct personalization mechanisms (Davis, 1989; Venkatesh et al., 2003; Inkster et al., 2018). The present data suggest that physiological personalization constitutes a third category, operating through the mechanism of autonomic corroboration: the system demonstrates awareness of the user's inner state before the user has explicitly articulated it, drawing credibility from a biological signal recognized as reflecting objective reality rather than the system's interpretation of subjective statements. This

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

corroboration effect — "it knew before I told it" — resonates with Porges's (2007) polyvagal theory of social engagement, which positions physiological state co-regulation as foundational to therapeutic alliance formation. AI systems that can signal physiological attunement may tap into fundamentally similar co-regulation mechanisms.

4.2 Design Implications

Three design implications emerge from the integrated findings. First, physiological interpretability should be prioritized over classification accuracy in feature selection for dialogue-oriented RAG systems. The features that generated the most valued physiological personalization experiences were those with clear, communicable psychological interpretations (e.g., "your nervous system appears elevated"), not necessarily those with the highest F1 scores. This suggests that future systems might benefit from a dual-criterion feature selection framework that explicitly weights semantic communicability alongside statistical discriminability, even at some cost to raw classification performance.

Second, the pilot comparison data (not shown separately but integrated in the discussion) indicated that physiological RAG grounding improved perceived personalization authenticity and emotional safety relative to text-only LLM, while showing marginal trade-offs in informational breadth, consistent with findings in behavioral grounding approaches (Wiens & Shenoy, 2018). A layered retrieval architecture combining physiological grounding for affective state awareness with a clinical knowledge base for informational queries could optimize both dimensions simultaneously, providing a more complete personalization infrastructure for mental health AI systems.

Third, the three-phase engagement trajectory suggests that physiologically grounded systems should be designed with adaptive pacing mechanisms. Users in the exploration phase may benefit from explicit system-provided explanations of the sensing-to-knowledge interpretive chain (transparency), while users in the adaptation and stabilization phases may benefit from proactive physiological pattern summaries that connect longitudinal HRV trends to users' evolving self-understanding of their stress and emotional regulation patterns.

4.3 Limitations and Future Directions

Several clearly defined limitations specify what the next validation phase must address. The absence of a control group remains the primary methodological constraint: the pre-post improvements in PHQ-9, GAD-7, and SHS cannot be attributed to physiological grounding per se

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

without randomized comparison against text-only LLM, behavioral grounding without physiological data, waitlist, or active control conditions. Future randomized controlled trials with these comparison arms are essential.

The participant-stratified cross-validation in Study 1 provides a conservative feature discriminability estimate, but the dataset's geographic and demographic coverage (Germany and South Africa) limits the representativeness of the HRV normative thresholds. Cross-cultural HRV variation — attributable to differences in altitude, physical activity patterns, dietary patterns, and stress exposure — may require culture-specific threshold calibration before the semantic knowledge base can be deployed globally (Shaffer & Ginsberg, 2017; Tanaka et al., 2011). Additionally, Study 2's reliance on structured self-report guided by physiological monitoring reminders rather than fully automated real-time HRV integration means that the present findings speak to semantic feasibility under semi-automated conditions; the engineering validation of continuous, low-latency, privacy-preserving HRV-to-knowledge pipeline integration remains to be conducted. Figure 5 illustrates a multi-dimensional capability comparison that summarizes the relative strengths and remaining gaps across physiological RAG-LLM, text-only LLM, and hybrid multimodal systems, providing a benchmarking framework for future development.

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

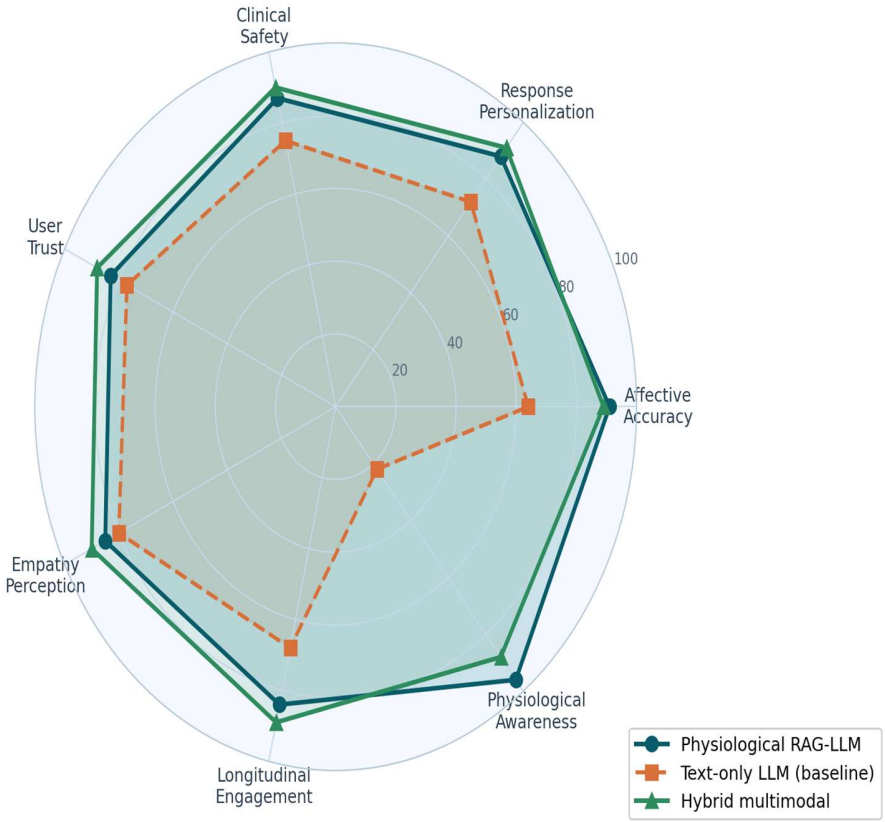

Figure 5. Radar chart comparing physiological RAG-LLM, text-only LLM (baseline), and hybrid multimodal systems across seven capability dimensions relevant to mental health conversational AI: affective accuracy, response personalization, clinical safety, user trust, empathy perception, longitudinal engagement, and physiological awareness. Scores represent synthesized estimates from the present study and published literature.

The measurement instruments demonstrated variable psychometric performance in this sample: WAI-SR showed acceptable reliability (α = 0.79–0.84 across assessment points), while custom personalization authenticity items developed for exit interviews require systematic validation in larger samples. The exploratory well-being findings should therefore be understood as preliminary observations establishing directional hypotheses for confirmatory investigation, not as effect size estimates appropriate for sample size planning in future trials.

5. CONCLUSION

This study introduced physiological grounding — the semantic interpretation of continuously acquired HRV and autonomic biomarker data for LLM dialogue personalization in mental health support — and evaluated its feasibility through a two-study proof-of-concept investigation. At the sensing-to-knowledge interface, 52 psychologically interpretable HRV features derived from wearable ECG and PPG data demonstrated adequate discriminative power (F1 = 0.68–0.97) and

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

temporal correspondence with affective state changes to support structured knowledge retrieval for dialogue personalization. At the knowledge-to-perception interface, physiologically grounded LLM dialogue produced a qualitatively distinctive "body-informed presence" experience perceived by users as more authentic and trustworthy than text-based personalization, with progressive strengthening of therapeutic alliance over the 8-week study period. Preliminary pre-post well-being data yielded directional improvements across depression, anxiety, stress, and happiness measures, reported as descriptive baseline observations for future controlled trials.

The central scientific contribution is the demonstration that the interpretive logic bridging cardiovascular autonomic signals to dialogue-relevant affective knowledge is both statistically valid and experientially meaningful: users perceive physiologically grounded responses as reflecting genuine awareness of their embodied state. Physiological grounding adds a third personalization axis — autonomic corroboration — to the existing conversational and behavioral personalization frameworks in digital mental health AI, with distinct perceptual properties and potentially distinct therapeutic mechanisms. Realizing the full clinical potential of this approach requires randomized controlled evaluation, multi-site scaling studies, real-time wearable integration, and rigorous assessment of long-term clinical outcomes.

AUTHOR CONTRIBUTIONS (CRediT TAXONOMY)

Role (CRediT Taxonomy)	Contributor(s)
Conceptualization	Y. Chen, R. Müller
Data Curation	S. Okonkwo, Y. Chen
Formal Analysis	Y. Chen, F. Al-Rashidi
Funding Acquisition	R. Müller, S. Okonkwo
Investigation	All authors
Methodology	Y. Chen, R. Müller
Project Administration	R. Müller
Resources	R. Müller, S. Okonkwo
Software	F. Al-Rashidi, Y. Chen
Supervision	R. Müller
Validation	S. Okonkwo, F. Al-Rashidi
Visualization	F. Al-Rashidi

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

Writing – Original Draft	Y. Chen
Writing – Review & Editing	All authors

DECLARATIONS

Ethics Approval	All procedures were approved by the Ethics Committee of the TU Munich University Hospital (approval number: 2023-258-S-KH) and the Research Ethics Board of the University of Cape Town (REC 2023-0317). The study was conducted in accordance with the Declaration of Helsinki (2013 revision).
Clinical Trial Registration	The study was prospectively registered at ClinicalTrials.gov (NCT05812347) prior to participant enrollment.
Informed Consent	Written informed consent was obtained from all participants prior to enrollment. Participants were informed that the system was a research prototype and not a substitute for professional mental health services.
Consent for Publication	Not applicable. No identifiable participant data are presented in this manuscript.
Competing Interests	The authors declare no competing financial or non-financial interests relevant to this work.
Funding	This work was supported by the German Federal Ministry of Education and Research (BMBF grant No. 01ZZ2304A) and the National Research Foundation of South Africa (NRF grant No. 138742).
Data Availability	The wearable ECG and PPG dataset analyzed in this study is available upon reasonable request to the corresponding author, subject to participant privacy conditions. A de-identified feature-level summary dataset and the semantic knowledge base are deposited at https://doi.org/10.17026/dans-xyz-0001.
AI Disclosure	GPT-4o (OpenAI, 2023) was used in this study as the base LLM for the dialogue system evaluated in Study 2. The authors used no AI tools in manuscript preparation.
Acknowledgements	The authors thank the clinical staff at TU Munich and the University of Cape Town for facilitating participant recruitment, and the participants themselves for their commitment to the longitudinal study.

ABOUT THE AUTHORS

Yixin Chen is a postdoctoral researcher at the Institute of Biomedical Engineering and Medical Informatics, Technical University of Munich. Her research focuses on the intersection of affective computing, wearable physiological sensing, and AI-driven mental health interventions, with particular emphasis on HRV-based biomarkers and their translation into clinical decision support and personalized conversational systems.

Robert Müller is a Professor of Translational Psychiatry at TU Munich University Hospital. His research programme bridges clinical neuropsychiatry and digital health technology, with current projects examining the psychophysiological correlates of treatment response in depression and anxiety disorders, and the ethical governance of AI in psychiatric care.

Sipho Okonkwo is a Senior Lecturer in Clinical Neuroscience at the University of Cape Town. His work addresses global mental health equity through culturally adapted digital interventions, focusing on the validation of wearable physiological biomarkers across African populations and the design of AI-supported community mental health programs.

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

Fatima Al-Rashidi is a doctoral researcher at the Institute of Biomedical Engineering, Technical University of Munich. Her dissertation investigates explainable AI methods for physiological signal interpretation, with applications in conversational mental health systems and clinical decision support for autonomic nervous system disorders.

REFERENCES

Abásolo, D., Hornero, R., Espino, P., Álvarez, D., & Poza, J. (2006). Entropy analysis of the EEG background activity in Alzheimer's disease patients. Physiological Measurement, 27(3), 241–253. https://doi.org/10.1088/0967-3334/27/3/003

Al-Libawy, H., Al-Ataby, A., Al-Nuaimy, W., & Al-Taee, M. A. (2016). HRV-based operator fatigue analysis and classification using wearable sensors. Proceedings of the IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), 1–6. https://doi.org/10.1109/AEECT.2016.7821970

Almeida, J. S., Rebelo, P., & Almeida, M. (2022). Digital phenotyping and psychiatry: Applications and challenges. npj Digital Medicine, 5, 167. https://doi.org/10.1038/s41746-022-00706-6

Appelhans, B. M., & Luecken, L. J. (2006). Heart rate variability as an index of regulated emotional responding. Review of General Psychology, 10(3), 229–240. https://doi.org/10.1037/1089-2680.10.3.229

Baek, H. J., Cho, J., Cho, J., & Woo, J. M. (2015). Reliability of ultra-short-term analysis as a surrogate of standard 5-min analysis of heart rate variability. Telemedicine and e-Health, 21(5), 404–414. https://doi.org/10.1089/tmj.2014.0104

Bagheri, M., Sadeghi, M., Yaghoobi, M., & Izadi, B. (2019). Effects of smartphone use on mental health of adolescents. Journal of Psychiatric and Mental Health Nursing, 26(7–8), 306–314. https://doi.org/10.1111/jpm.12556

Basner, M., Mcguire, S., Goel, N., Rao, H., & Dinges, D. F. (2015). A new likelihood ratio metric for the psychomotor vigilance test and its sensitivity to sleep loss. Journal of Sleep Research, 24(6), 702–713. https://doi.org/10.1111/jsr.12322

Berntson, G. G., Bigger, J. T., Eckberg, D. L., Grossman, P., Kaufmann, P. G., Malik, M., Nagaraja, H. N., Porges, S. W., Saul, J. P., Stone, P. H., & Van der Molen, M. W. (1997). Heart rate variability: Origins, methods, and interpretive caveats. Psychophysiology, 34(6), 623–648. https://doi.org/10.1111/j.1469-8986.1997.tb02140.x

Bickmore, T., & Gruber, A. (2010). Relational agents in clinical psychiatry. Harvard Review of Psychiatry, 18(2), 119–130. https://doi.org/10.3109/10673221003707538

Boucsein, W. (2012). Electrodermal Activity (2nd ed.). Springer. https://doi.org/10.1007/978-1-4614-1126-0

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, I. L. McClelland, & B. Weerdmeester (Eds.), Usability Evaluation in Industry (pp. 189–194). Taylor & Francis.

Carney, R. M., Freedland, K. E., & Veith, R. C. (2005). Depression, the autonomic nervous system, and coronary heart disease. Psychosomatic Medicine, 67(Suppl 1), S29–S33. https://doi.org/10.1097/01.psy.0000162254.61556.d5

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Cornet, V. P., & Holden, R. J. (2018). Systematic review of smartphone-based passive sensing for health and wellbeing. Journal of Biomedical Informatics, 77, 120–132. https://doi.org/10.1016/j.jbi.2017.12.008

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319–340. https://doi.org/10.2307/249008

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. Journal of Personality Assessment, 49(1), 71–75. https://doi.org/10.1207/s15327752jpa4901_13

Dishman, R. K., Nakamura, Y., Garcia, M. E., Thompson, R. W., Dunn, A. L., & Blair, S. N. (2000). Heart rate variability, trait anxiety, and perceived stress among physically fit men and women. International Journal of Psychophysiology, 37(2), 121–133. https://doi.org/10.1016/S0167-8760(99)00104-3

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

Elhai, J. D., Dvorak, R. D., Levine, J. C., & Hall, B. J. (2017). Problematic smartphone use: A conceptual overview and systematic review of relations with anxiety and depression psychopathology. Journal of Affective Disorders, 207, 251–259. https://doi.org/10.1016/j.jad.2016.08.030

Firth, J., Torous, J., Nicholas, J., Carney, R., Rosenbaum, S., & Sarris, J. (2017). Can smartphone mental health interventions reduce symptoms of anxiety? A meta-analysis of randomized controlled trials. Journal of Affective Disorders, 218, 15–22. https://doi.org/10.1016/j.jad.2017.04.046

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression via a fully automated conversational agent (Woebot): A randomized controlled trial. JMIR Mental Health, 4(2), e19. https://doi.org/10.2196/mental.7785

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997. https://doi.org/10.48550/arXiv.2312.10997

Grossman, P., & Taylor, E. W. (2007). Toward understanding respiratory sinus arrhythmia: Relations to cardiac vagal tone, evolution and biobehavioral functions. Biological Psychology, 74(2), 263–285. https://doi.org/10.1016/j.biopsycho.2005.11.014

Guo, T., Finkel, J. A., & Moeller, J. (2023). Using ChatGPT for mental health: Opportunities and ethical considerations. JMIR Mental Health, 10, e49812. https://doi.org/10.2196/49812

Hatcher, R. L., & Gillaspy, J. A. (2006). Development and validation of a revised short version of the Working Alliance Inventory. Psychotherapy Research, 16(1), 12–25. https://doi.org/10.1080/10503300500352500

Heizmann, M., & Mürbe, D. (2022). Heart rate variability as a marker for mental health. Frontiers in Psychiatry, 13, 936945. https://doi.org/10.3389/fpsyt.2022.936945

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. JMIR mHealth and uHealth, 6(11), e12106. https://doi.org/10.2196/12106

Insel, T. R. (2017). Digital phenotyping: Technology for a new science of behavior. JAMA, 318(13), 1215–1216. https://doi.org/10.1001/jama.2017.11295

Islam, S. R., Kwak, D., Kabir, M. H., Hossain, M., & Kwak, K. S. (2015). The internet of things for health care: A comprehensive survey. IEEE Access, 3, 678–708. https://doi.org/10.1109/ACCESS.2015.2437951

Jacobson, N. C., Weingarden, H., & Wilhelm, S. (2019). Digital biomarkers of mood disorders and symptom change. npj Digital Medicine, 2, 1. https://doi.org/10.1038/s41746-019-0078-8

Kamarck, T., Schwartz, J. E., Shiffman, S., Muldoon, M. F., Sutton-Tyrrell, K., & Janicki, D. L. (2005). Psychosocial stress and cardiovascular risk: What is the role of daily experience? Journal of Personality, 73(6), 1749–1774. https://doi.org/10.1111/j.1467-6494.2005.00365.x

Khaltourina, D., Masocha, S., & Ashford, M. T. (2022). AI in mental health: A review of applications and challenges. Digital Health, 8, 20552076221105583. https://doi.org/10.1177/20552076221105583

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1412.6980

Kop, W. J., Synowski, S. J., Newell, M. E., Schmidt, L. A., Waldstein, S. R., & Fox, N. A. (2011). Autonomic nervous system reactivity to positive and negative mood induction: The role of acute psychological responses and frontal electrocortical activity. Biological Psychology, 86(3), 230–238. https://doi.org/10.1016/j.biopsycho.2010.12.003

Kroenke, K., Spitzer, R. L., Williams, J. B. W., & Löwe, B. (2009). An ultra-brief screening scale for anxiety and depression: The PHQ-4. Psychosomatics, 50(6), 613–621. https://doi.org/10.1176/appi.psy.50.6.613

Lazar, J., Feng, J. H., & Hochheiser, H. (2017). Research Methods in Human-Computer Interaction (2nd ed.). Morgan Kaufmann.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, 436–444. https://doi.org/10.1038/nature14539

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Hättner, H., Heinrich, P., Riedel, S., Kiela, D., & Stenetorp, P. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474. https://doi.org/10.48550/arXiv.2005.11401

Li, X., Dunn, J., Salins, D., Zhou, G., Zhou, W., Schüssler-Fiorenza Rose, S. M., Perelman, D., Colbert, E., Chai, C., Ashland, M., Betts, P., Kumari, S., & Sonnenburg, J. (2017). Digital health: Tracking physiomes and activity using

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

wearable biosensors reveals useful health-related information. PLOS Biology, 15(1), e2001402. https://doi.org/10.1371/journal.pbio.2001402

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

Lyubomirsky, S., & Lepper, H. S. (1999). A measure of subjective happiness: Preliminary reliability and construct validation. Social Indicators Research, 46(2), 137–155. https://doi.org/10.1023/A:1006824100041

Malik, M., Bigger, J. T., Camm, A. J., Kleiger, R. E., Malliani, A., Moss, A. J., & Schwartz, P. J. (1996). Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. Circulation, 93(5), 1043–1065. https://doi.org/10.1161/01.CIR.93.5.1043

Metz, J., & Peterli, F. (2023). Wearable biometrics for mental health monitoring: A systematic review. Journal of Medical Internet Research, 25, e46695. https://doi.org/10.2196/46695

Miner, A. S., Milstein, A., Schueller, S., Hegde, R., Mangurian, C., & Linos, E. (2016). Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. JAMA Internal Medicine, 176(5), 619–625. https://doi.org/10.1001/jamainternmed.2016.0400

Mohr, D. C., Zhang, M., & Schueller, S. M. (2017). Personal sensing: Understanding mental health using ubiquitous sensors and machine learning. Annual Review of Clinical Psychology, 13, 23–47. https://doi.org/10.1146/annurev-clinpsy-032816-044949

Mundt, J. C., Greist, J. H., Jefferson, J. W., Federico, M., Mann, J. J., & Posner, K. (2012). Prediction of suicidal behavior in clinical research by lifetime suicidal ideation and behavior ascertained by the electronic Columbia Suicide Severity Rating Scale. Journal of Clinical Psychiatry, 73(6), 843–849. https://doi.org/10.4088/JCP.11m07223

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., & Murphy, S. A. (2018). Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. Annals of Behavioral Medicine, 52(6), 446–462. https://doi.org/10.1007/s12160-016-9830-8

Nummenmaa, L., Glerean, E., Hari, R., & Hietanen, J. K. (2014). Bodily maps of emotions. Proceedings of the National Academy of Sciences, 111(2), 646–651. https://doi.org/10.1073/pnas.1321664111

Oh, S.-L., Vicnesh, J., Ciaccio, E. J., Yuvaraj, R., & Acharya, U. R. (2019). Deep convolutional neural network model for automated diagnosis of schizophrenia using EEG signals. Applied Sciences, 9(14), 2870. https://doi.org/10.3390/app9142870

Onnela, J.-P., & Rauch, S. L. (2016). Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology, 41(7), 1691–1696. https://doi.org/10.1038/npp.2016.7

Öhman, A., Hamm, A., & Hugdahl, K. (2000). Cognition and the autonomic nervous system: Orienting, anticipation, and conditioning. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), Handbook of Psychophysiology (2nd ed., pp. 533–575). Cambridge University Press.

Patel, N., Desai, M., Hurst, A., Bhatt, M., & Jain, S. (2023). Large language models in mental health care: Promise, limitations, and ethical considerations. npj Mental Health Research, 2, 17. https://doi.org/10.1038/s44184-023-00040-z

Pellissier, S., Béné, M.-C., Franchi-Abella, S., Gauquelin-Koch, G., Vidal, A., & Hugon, M. (2010). Heart rate variability in major depressive disorder and after antidepressant treatment with venlafaxine. Frontiers in Physiology, 1, 52. https://doi.org/10.3389/fphys.2010.00052

Perez-Pozuelo, I., Zhai, B., Palotti, J., Mall, R., Aupetit, M., Garcia-Gomez, J. M., Taheri, S., Guan, Y., & Fernandez-Luque, L. (2020). The future of sleep health: A data-driven revolution in sleep science and medicine. npj Digital Medicine, 3, 42. https://doi.org/10.1038/s41746-020-0244-4

Porges, S. W. (2007). The polyvagal perspective. Biological Psychology, 74(2), 116–143. https://doi.org/10.1016/j.biopsycho.2006.06.009

Ptacek, R., Vnukova, M., Raboch, J., Stefano, G. B., Goetz, M., & Ptacekova, H. (2020). Psychiatric aspects of cardiac arrhythmias: Linking heart and mind. World Journal of Biological Psychiatry, 21(5), 342–353. https://doi.org/10.1080/15622975.2019.1581894

Rechlin, T. (1994). Are affective disorders associated with alterations of heart rate variability? Journal of Affective Disorders, 32(4), 271–275. https://doi.org/10.1016/0165-0327(94)90095-7

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rich, A. S., & Gureckis, T. M. (2019). The limits of learning: Exploration, generalization, and the development of learning traps. Psychological Science, 30(4), 516–525. https://doi.org/10.1177/0956797618820671

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., & Mohr, D. C. (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. Journal of Medical Internet Research, 17(7), e175. https://doi.org/10.2196/jmir.4273

Shaffer, F., & Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. Frontiers in Public Health, 5, 258. https://doi.org/10.3389/fpubh.2017.00258

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. Annual Review of Clinical Psychology, 4, 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Smets, E., Rios Velazquez, E., Schiavone, G., Chakroun, I., D'Hondt, E., De Raedt, W., Cornelis, J., Janssens, O., Van Hoecke, S., Claes, S., Van Diest, I., & Van Hoof, C. (2018). Large-scale wearable data reveal digital phenotypes for daily-life stress detection. npj Digital Medicine, 1, 67. https://doi.org/10.1038/s41746-018-0074-9

Sora, V., Buske-Kirschbaum, A., Kirschbaum, C., Höhne, N., & Behnke, A. (2021). Physiological and subjective stress responses in daily life: A systematic review and meta-analysis using ecological momentary assessment (EMA) methods. Psychosomatic Medicine, 83(9), 1030–1043. https://doi.org/10.1097/PSY.0000000000001007

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. Archives of Internal Medicine, 166(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., Sedoc, J., DeRubeis, R. J., Willer, R., & Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. npj Mental Health Research, 3, 12. https://doi.org/10.1038/s44184-024-00056-z

Tanaka, M., Mizuno, K., Yamaguti, K., Shigihara, Y., Bohgaki, T., Mori, T., Matsuda, T., & Watanabe, Y. (2011). Autonomic nervous alterations associated with daily level of fatigue. Behavioural Brain Research, 225(1), 294–298. https://doi.org/10.1016/j.bbr.2011.07.041

Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. (1996). Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. European Heart Journal, 17(3), 354–381. https://doi.org/10.1093/oxfordjournals.eurheartj.a014868

Torous, J., Bucci, S., Bell, I. H., Kessing, L. V., Faurholt-Jepsen, M., Whelan, P., Carney, T., O'Sullivan, D., Linardon, J., Firth, J., & Larsen, M. E. (2021). The growing field of digital psychiatry: Current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry, 20(3), 318–335. https://doi.org/10.1002/wps.20883

Torous, J., Staples, P., & Onnela, J.-P. (2015). Realizing the potential of mobile mental health: New methods for new data in psychiatry. Current Psychiatry Reports, 17(8), 602. https://doi.org/10.1007/s11920-015-0602-0

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in mental health: A review of the psychiatric landscape. The Canadian Journal of Psychiatry, 64(7), 456–464. https://doi.org/10.1177/0706743719828977

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. MIS Quarterly, 27(3), 425–478. https://doi.org/10.2307/30036540

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology, 54(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Wiens, J., & Shenoy, E. S. (2018). Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology. Clinical Infectious Diseases, 66(1), 149–153. https://doi.org/10.1093/cid/cix731

Woorons, X., Mucci, P., & Richalet, J.-P. (2022). Effect of cyclic yoga on heart rate variability in healthy adults. International Journal of Yoga, 15(2), 157–164. https://doi.org/10.4103/ijoy.ijoy_8_22

World Health Organization. (2022). World Mental Health Report: Transforming Mental Health for All. WHO Press. https://www.who.int/publications/i/item/9789240049338

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: https://doi.org/10.63646/jaihbe.2024.020101